PAPER V1.3 · NOVEMBER 2026

The Panel Is the Computer

Infinite Sunlight. Effortless Intelligence.

Paper v1.3 · Public wording revision, 8 November 2026. Clarifies the infrastructure-only scope, measurement boundaries and public disclosure level. The v1.3 technical assumptions and historical revision notes are retained; this is not a new validation result.

We measure success by how efficiently we execute existing AI models and deliver their output. We are developing the infrastructure, while model creators and users determine what that output accomplishes. Vela's distributed architecture is designed to maximize token throughput per unit of sunlight, launched mass, and deployed cost.

Summary. Project Vela is developing infrastructure to execute existing AI models. Its orbital design integrates compute with solar power and passive heat rejection. The third-generation design target is about **5 billion tokens per second processed from one 100-tonne Starship launch** on a 7–8B-class language model (input and output combined under an input-heavy reference scenario, not split-independent), using ~75 kW per satellite; ~4–5 MW per Starship launch today, ~6 MW with our next-generation panels. Earlier v1.3 calculations expressed this target as roughly **100,000 H100 GPUs**. That is a source-derived illustrative calculation, not an established matched-workload GPU comparison. The approximately billion-token-per-second launch figure is a calculated projection from energy per operation measured while a chip ran an image model, applied to a language-model workload—distinct from the language-model bench results in §1.6; it is not measured language-model throughput or delivered launch capacity. The current V1 rating is a measurement-based feed-forward compute projection, explained in §1.3; earlier solar-watt and GPU comparisons are retained as historical calculations. This paper preserves the numerical assumptions and history while distinguishing measurements, calculations and design targets.

1.1 The problem: AI runs on power

A 100,000-GPU data center draws well over 100 megawatts once servers and cooling are counted. New sites wait years for grid connections, and they need land and water for cooling. Chips are no longer the scarcest input to AI. Power and permission are.

1.2 The idea: put the computer where the power is

Above the atmosphere, sunlight delivers about 1,361 watts per square metre, and in the right orbit it shines nearly all the time. Vela's design integrates inference compute with the solar array. Passive heat rejection is a design objective; thermal performance and environmental durability require validation.

The basic unit is a **compute panel** of about 2.4 × 4.35 m, delivering **~3.8 kW per panel (nameplate), about 3 kW in orbit**. It's made of tiles that each carry their own solar cell and compute. Spare tiles are built in, so a failed tile is switched out, not repaired. A satellite carries many panels.

1.3 Measurement-based compute projection

V1 / MEASUREMENT-BASED PROJECTION

~50 token positions/s per accelerator chip*

* Qwen2.5-7B feed-forward compute projection from batched lab measurements; excludes attention and system overhead. Not measured output generation.

Calculation basis →

Calculation basis. For Qwen2.5-7B, measured batched processing of a representative feed-forward component normalizes to about 51 full-model feed-forward-equivalent token positions/s per allocated accelerator chip, rounded to 50. This assumes comparable service across layers, sufficient model residency and sustained batching. Attention, memory handling, transfers and scheduling are outside this compute rating. It is not single-user generation speed or complete-system throughput. No power assumption is used in this rating.

Historical solar-efficiency calculations. The following v1.3 table and H100 comparison preserve earlier assumptions; they are not the current V1 performance rating. V2/V3 entries remain design targets.

Our benchmark is token throughput per watt of electrical power delivered by the solar cells, for a stated 7–8B-class language model—not per watt of incident sunlight. V1 is calculated from measured chip throughput; V2 and V3 are design targets. Chip, module and complete-system boundaries must be distinguished. Token counts are not universally comparable across models, tokenizers, context, precision or input/output mix. Fidelity is checked against the stated baseline. Launched-mass and deployed-cost efficiency remain design objectives.

Generation	Status	What changes	Tokens/s per solar watt	vs. an NVIDIA H100, per watt
V1	**Building now.** Chip speed measured on the bench	Off-the-shelf low-power AI chips on a solar panel	14–18 (calculated)	**Roughly on par** (provisional calculation; 0.7–0.9× on the stated basis)
V2	**In development** (design target)	A custom chip designed for this one job	200–370 (target)	**~10–18×**
V3	**In design**	The panel itself becomes the computer	650–1,300 (target)	**~30–65×**

Unit convention: "tok/W" in the calculation tables means tokens per second per watt (tokens per joule). Solar watts are electrical power delivered by the solar cells, not incident sunlight. Reading context and generating output have different costs; all comparisons depend on the stated model and workload assumptions.

Source-derived H100 calculation, not a matched-workload comparison. The v1.3 calculation uses NVIDIA's published MLPerf Inference results: eight H100s serving Llama 2 70B at about 2.1–2.4 tokens per second per watt of server power, then scales that figure to an 8B-class model at roughly 20 tokens per second per watt. V1's calculated 14–18 is 0.7–0.9× that estimate; we describe this as roughly on par: a provisional calculated comparison based on the stated historical bench data and system assumptions. Measured bench configurations have differed in per-chip throughput, and the comparison will be revisited as measurements continue. The different model sizes, workload assumptions and power-accounting boundaries do not establish matched-workload parity. Vela's denominator is electrical power delivered by the solar cells; the GPU denominator is server power, excluding data-center cooling. These calculations are retained for transparency, not presented as a demonstrated efficiency comparison.

Two different calculations. The V1 14–18 tokens/s per solar watt above is the original v1.3 calculated case, and the H100 comparison is tied to that case on the stated basis; it is not an upgraded current ratio. The later historical input-processing panel scenario in §1.6 (about 46,000 tokens/s at 7–8B class per ~3.8 kW nameplate, roughly 12 per watt) is a separate calculation with newer, different assumptions. The two are not simultaneous validated results; neither is the current V1 compute rating.

1.4 What one launch is designed to do

- **~5 billion tokens per second processed*** on a 7–8B-class model, in sunlight
- **≈ 100,000 H100 GPUs** in the historical source-derived calculation; not an established matched-workload comparison
- **~75 kW per satellite; ~4–5 MW per Starship launch today, ~6 MW with our next-generation panels** (~25,000 m² of panel at ~240 W/m², in sunlight, for the ~6 MW next-generation-panel case)
- **80 satellites** per 100-tonne Starship launch
- **Purpose-built silicon:** a design target, not flight-qualified hardware

* Design target for purpose-built silicon, not a measurement. "Processed" counts input and output tokens together under an input-heavy reference scenario; that scenario is a reference point, not a chosen operating split. Output generation costs more per token, so the total depends on the input/output mix. Power basis: electrical power delivered by the solar cells, in sunlight—not incident sunlight.

Separately, a calculated projection of about a billion tokens per second per launch applies energy per operation measured while a chip ran an image model to a 7–8B-class language model. It is distinct from the language-model bench results,

which measured accuracy and layer throughput, not energy per token. Language-model energy measurements are under way. Our purpose-built silicon targets five billion processed; neither figure is delivered launch capacity.

The arithmetic, rung by rung, so you can check it:

Step	Value	Basis
Panel area per launch	~25,000 m²	80 satellites × ~314 m² each
Solar power per m²	~240 W	Silicon space cells at ~17.6% mid-life (published space-agency data), × 1,361 W/m²
Solar power per launch	**~4–5 MW today; ~6 MW with next-generation panels**	~75 kW per satellite; ~6 MW is 25,000 m² × 240 W/m² with next-generation panels
Historical V1 launch calculation	~0.1B tok/s	6 MW (next-generation panels) × 14–18 tok/W in sunlight — off-the-shelf chips; calculated from chip speed measured on our bench
At measured shipping-chip efficiency	~0.9B tok/s	6 MW (next-generation panels) × ~150 tok/W in sunlight — shipping low-power silicon measured on our bench (vision workload), projected to a 7–8B LLM
Custom chip at published ratings (V2 target)	1.2–2.2B tok/s	6 MW (next-generation panels) × 200–370 tok/W in sunlight — published chip ratings assuming high utilization (not yet shown)
Purpose-built panel silicon (V3 target)	**~5 billion tok/s**	**6 MW (next-generation panels) × ~850 tok/W** in sunlight (V3 design target; the 650–1,300 band depends on workload and utilization assumptions)

The original v1.3 MLPerf-based calculation puts the 5-billion-token-per-second design target closer to 200,000 H100s; the separate peak-math estimate, using assumed utilization, gives 75,000–125,000. The earlier headline used 100,000. These differing source-derived calculations are retained as historical assumptions, not reconciled into a new result and not evidence of matched-workload equivalence.

The V2 and V3 rungs are design targets, quoted per 100-tonne launch and in sunlight. Nothing at this scale has been built or flown yet.

Why the rungs differ. The billion-token projection applies chip energy per operation measured on an image model to a 7–8B-class language model. It is not a language-model throughput measurement. V1's lower calculated figure includes off-the-shelf system overhead. The purpose-built panel and silicon figures remain design targets.

1.5 Design rationale and validation

Building it is hard engineering work, but nothing in the plan is waiting on an invention.

1. In orbit, power is the free resource. On the ground, AI is limited by how much grid power a site can get. In orbit, the limit is how much panel you launch. Silicon solar cells are among the most mass-produced things on Earth: the world makes hundreds of gigawatts of them a year. One launch needs about a thousandth of a percent of that.

2. Efficient distributed execution. Energy efficiency depends on the complete inference system, not arithmetic alone. Vela is developing a novel distributed architecture to execute existing models. Its efficiency targets remain subject to hardware and whole-system validation.

3. Passive thermal design. The orbital design uses passive heat rejection as a design objective. Thermal performance and environmental durability require validation; no flight-qualified system is claimed.

4. Manufacturing and deployment validation. Manufacturing and deployment require validation; no flight-qualified system is claimed.

1.6 Measured so far

Measured on development hardware, not modeled. November 2026.

- **Complete-model accuracy, through 3B.** Complete 0.5B, 1.5B and 3B language models were evaluated in a hybrid system: the accelerators executed part of each model and a host computer executed the rest at full precision. Perplexity rose by +0.85% (0.5B), +0.77% (1.5B) and +0.54% to +0.96% (3B, depending on configuration) relative to the full-precision reference model on the same evaluation text. Configurations differ between models. Subsequent 0.5B

reruns showed a +0.79% to +0.84% perplexity increase. These runs measured correctness, not sustained speed. They are complete-model results in a hybrid system, distinct from a model running entirely on the accelerators.

- **7B hybrid accuracy and 72B samples.** The 7B model showed a 0.55% perplexity increase against full precision with all feed-forward layers on the accelerator and the remaining operations on the host. Sampled 72B layers closely matched full-precision outputs. 7B: complete hybrid-model accuracy, not serving speed. 72B: sampled layers only; no whole-model perplexity result.
- **Parallel throughput.** Four parallel compute units delivered about 4× the aggregate layer-processing throughput of one unit. Layer processing only; this does not establish whole-model throughput scaling.
- **Batching fidelity checked.** In the tested configurations, batched execution produced bit-for-bit identical answers. This result does not establish the behavior of every model or configuration.
- **Power characterization.** Additional language-layer power was below the wall meter's resolution. Power at the throughput benchmark conditions is not yet established; direct module metering is next. The earlier 1.5 W figure is a planning assumption, not a measured maximum.

Calculated panel scenarios (input processing). About 105,000 tokens per second at 3B class, or about 46,000 at 7–8B class, per ~3.8 kW (nameplate) panel. Calculated from layer throughput measured on our bench—not measured on a panel. Assumes accelerator power fits the panel budget, the controller keeps pace and work stays fully batched; none of these is yet verified. These panel scenarios cover input processing only; they are not measured end-to-end serving rates.

1.7 The path: three generations, three programs at once

We're running all three generations in parallel rather than one after another. That's how we intend to move fast.

- **V1 is intended to validate the system:** real models, real accuracy, real power, built from parts anyone can buy. Its per-watt comparison is a source-derived calculation, not established matched-workload parity. Our bench is now measuring a whole system to confirm it.
- **V2: purpose-built silicon.** Improved efficiency for executing existing models is a design target, subject to hardware and whole-system validation.
- **V3: the panel itself becomes the computer.** Panel integration and efficiency are design targets that require validation.

1.8 Even if we're wrong by 100×

The V1 calculation works out to **tens of millions of tokens per second per launch** under the published assumptions. This is a design proposal, not demonstrated orbital capacity. System performance, communications, environmental durability and operational infrastructure remain to be validated, even without the targeted V2 and V3 efficiency gains.

1.9 What still has to be proven

We'd rather say this plainly:

- **Custom-chip efficiency on real language models.** Datasheet figures overstate what chips deliver. Measurements, calculations and design targets are reported separately.
- **High chip utilization on language models.** Off-the-shelf chips deliver only a fraction of their rated throughput on language models; our V2 and V3 targets assume high utilization, which is the first claim to prove.
- **Input-reading vs. answer-writing mix.** The ~850 tok/W figure assumes workloads weighted toward reading context; generating output tokens alone runs lower depending on memory design.
- **Building the third-generation panel at high yield.** Each manufacturing step exists in volume today; combining them at panel scale hasn't been done yet.
- **Solar cell output over years in orbit, at operating temperature.** We plan with conservative, published space-agency figures.
- **Launch at full rated payload.** Our figures are per 100-tonne launch, Starship's rated capacity. Flights so far have carried less.
- **Ground links and customers** at this scale.

Each item has a test or a study behind it.

1.10 Who we are

Project Vela is a product of Alien Alliances, Inc., based in Orem, Utah, and is led by its co-founders: Jason Crawford, Founder & CEO, and Jennifer Sharp, Co-Founder, CFO & Head of Quality. Public company information and bios are on About; collaboration inquiries can be directed to info@project-vela.com.

Want the engineering? We share architecture, test data and the full physics and manufacturing case with researchers, partners and investors under NDA: **info@project-vela.com**. Follow progress at **@ProjectVela** on X.

1.11 Notes

- *Forward-looking figures.* V2 and V3 figures are design targets, not measurements. §1.6 reports development-hardware measurements, not a whole-system power measurement. V1's per-watt figure is calculated from chip speed measured there; a whole-system measurement is under way. Nothing described here has flown.
- *Revision.* v1.1 (4 Nov 2026): V1's per-watt figure is now labelled as calculated from measured chip speed; v1 described it as measured. §1.6 now reports the silicon results at 0.5B and 1.5B (within about 1%) in place of v1's software-simulation figure for larger models (0.1–0.4%), which the chips did not bear out, and adds the multi-module speed result. The power line now reads "under ~1.5 W per chip, a ceiling" (v1: "under ~1 W"): v1's bound covered only the extra power while running, not the chip's own idle draw.
- *Revision.* v1.2 (5 Nov 2026): adds the billion-token figure for shipping chips (summary, §1.4).
- *Revision.* v1.3 (5 Nov 2026): §1.4 shows the launch arithmetic as a ladder from today's bench to the V3 target; §1.9 adds chip utilization and the reading/writing mix; model class unified as 7–8B.
- *Public wording revision* (6 Nov 2026, evening): §1.6 expands bench results (complete-model accuracy through 3B in a hybrid accelerator-and-host system; sampled 7B and 72B layers; parallel layer throughput) and adds calculated input-processing panel scenarios; the 5-billion figure is labeled as tokens processed under an input-heavy reference scenario; the H100 comparison is labeled a provisional calculation. Not a new scientific version.
- *Billion-token basis:* energy per operation measured on our bench while the chip ran an image model, applied to a 7–8B-class language model at ~15 billion operations per token, with ~85% of solar power reaching the chips (a design target). Language-model energy measurements are under way.
- *H100 basis:* MLPerf Inference results for 8 × H100 on Llama 2 70B, at server power, scaled to 8B class.
- *Solar basis:* AM0 sunlight of 1,361 W/m². Silicon space-cell efficiency from published CNES/CEA results (2025).
- *Energy floor:* the thermodynamic minimum energy to erase one bit (Landauer limit) at panel temperature.

> We measure success by how efficiently we execute existing AI models and deliver their output. We are developing the infrastructure, while model creators and users determine what that output accomplishes. Vela's distributed architecture is designed to maximize token throughput per unit of sunlight, launched mass, and deployed cost.
>
> · How we get there →